Scalable Systems Software for Terascale Computer Centers

Al Geist
Oak Ridge National Lab

SciDAC PI Meeting
March 10, 2003

Research sponsored by MICS Office of DOE

Scalable Systems Software for Terascale Computer Centers

Coordinator: Al Geist

Participating Organizations

Includes DOE Labs, NSF Supercomputer Centers, Vendors

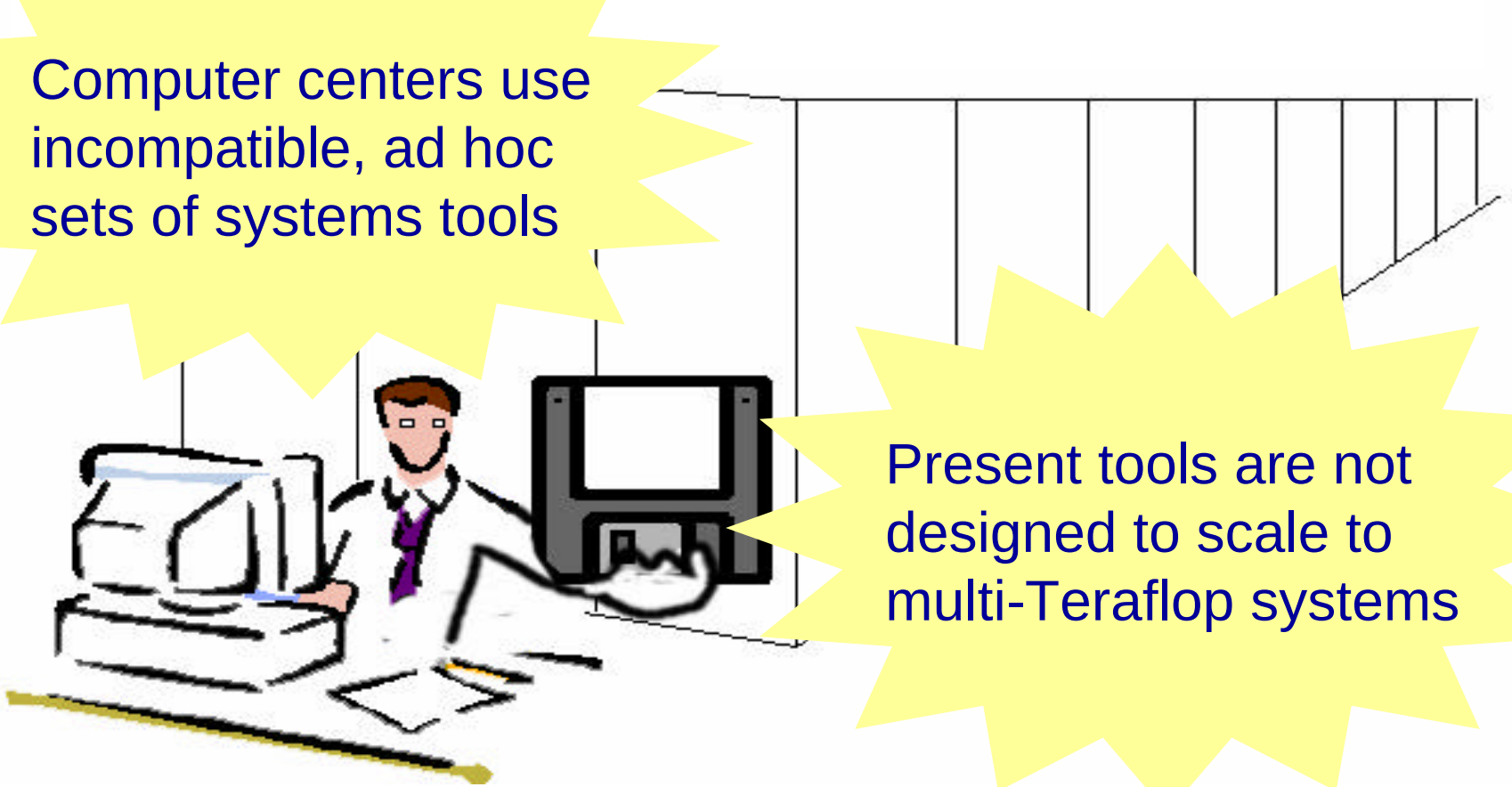
ORNL	**Ames**	**NCSA**	**Cray**
ANL	**SNL**	**SDSC**	**Intel**
LBNL	**LANL**	**IBM**	**Unlimited Scale**
PNNL	**PSC**	**Open to all like MPI forum**	
Clemson			

www.scidac.org/ScalableSystems

The Problem Today

System administrators and managers of terascale computer centers are facing a crisis:

Computer centers use incompatible, ad hoc sets of systems tools



Present tools are not designed to scale to multi-Teraflop systems

Three Goals

Collectively (with industry) agree on and specify standardized interfaces between system components

MPI-like process to promote interoperability, portability, and long-term usability.

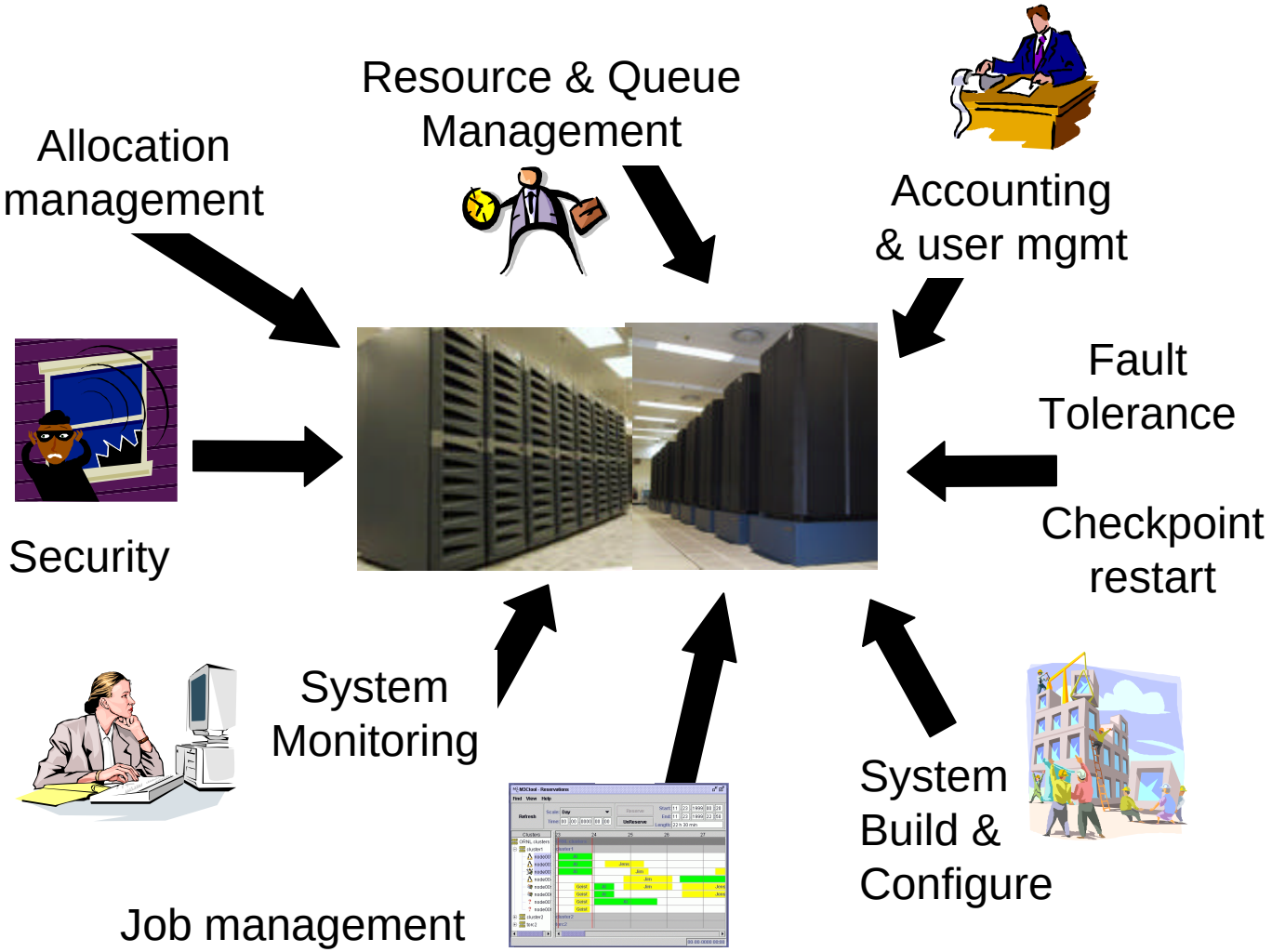
Produce a fully integrated suite of systems software and tools

Reference Implementation for the management and utilization of terascale computational resources.

Research and development of more advanced versions of the components

To support the scalability, fault tolerance, and performance requirements of large science applications. Up to 10,000 nodes.

Scope of the Effort



Impact

Fundamentally change the way future high-end systems software is developed and distributed

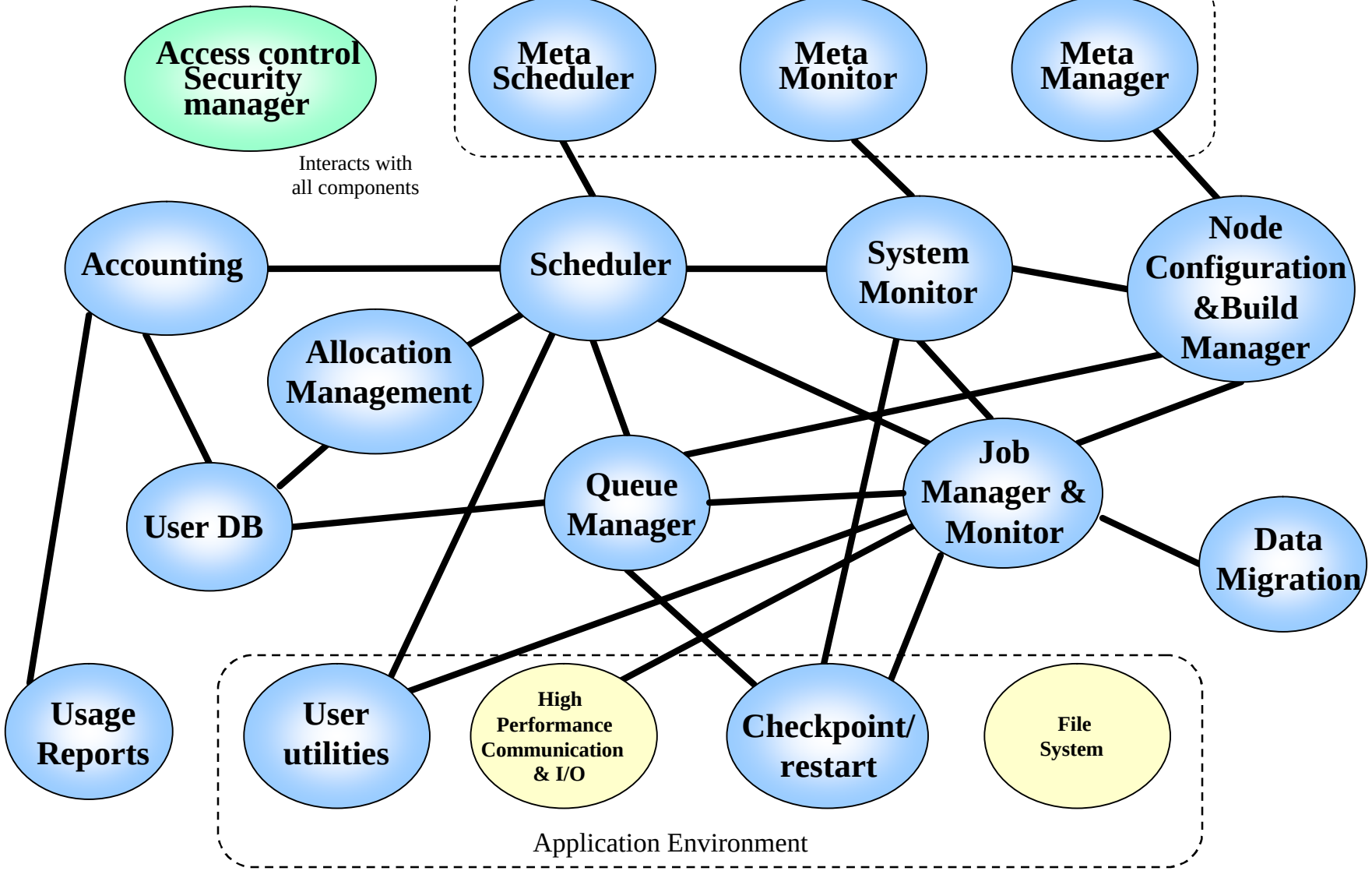
Reduced facility management costs

- reduce need to support ad hoc software
- better systems tools available
- able to get machines up and running faster and keep running

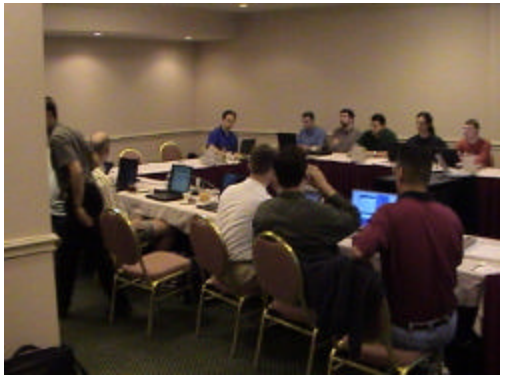
More effective use of machines by scientific applications

- scalable launch of jobs and checkpoint/restart
- job monitoring and management tools
- allocation management interface

System Software Architecture



Project Management



Quarterly Face-to Face Meetings

To discuss and vote on interface proposals

Four different Working Groups

1. Node build, configuration, and information service
2. Resource management, scheduling, and allocation
3. Proccess management, system monitoring, and checkpointing
4. Validation and Integration

Web-based Project Notebooks (over 200 pages and growing)

A main notebook for general information & mtg notes
And individual notebooks for each working group

www.scidac.org/ScalableSystems

System Software Components

Presently under construction

Strong Emphasis on multi-lab cooperation and team effort

Meta Scheduler *D. Jackson*

Scheduler *D. Jackson*

Accounting *S. Jackson*

Allocation Management *S. Jackson*

Queue Manager *B. Bode*

Resource Mgmt Working Group

API Tester Validator *W. McClendon*

Meta Manager *S. Scott*

Node Manager *T. Naughton*

Package Services *J. Mugler*

Service Directory *N. Desai*

Information Services *J.P. Navarro*

Authentication & Communication *N. Desai, A. Lusk*

Build & Configure Working Group

System/Job Monitors *M. Showerman*

Job Manager *B. Bode*

C-Plant XML interface *E. Debenedictis*

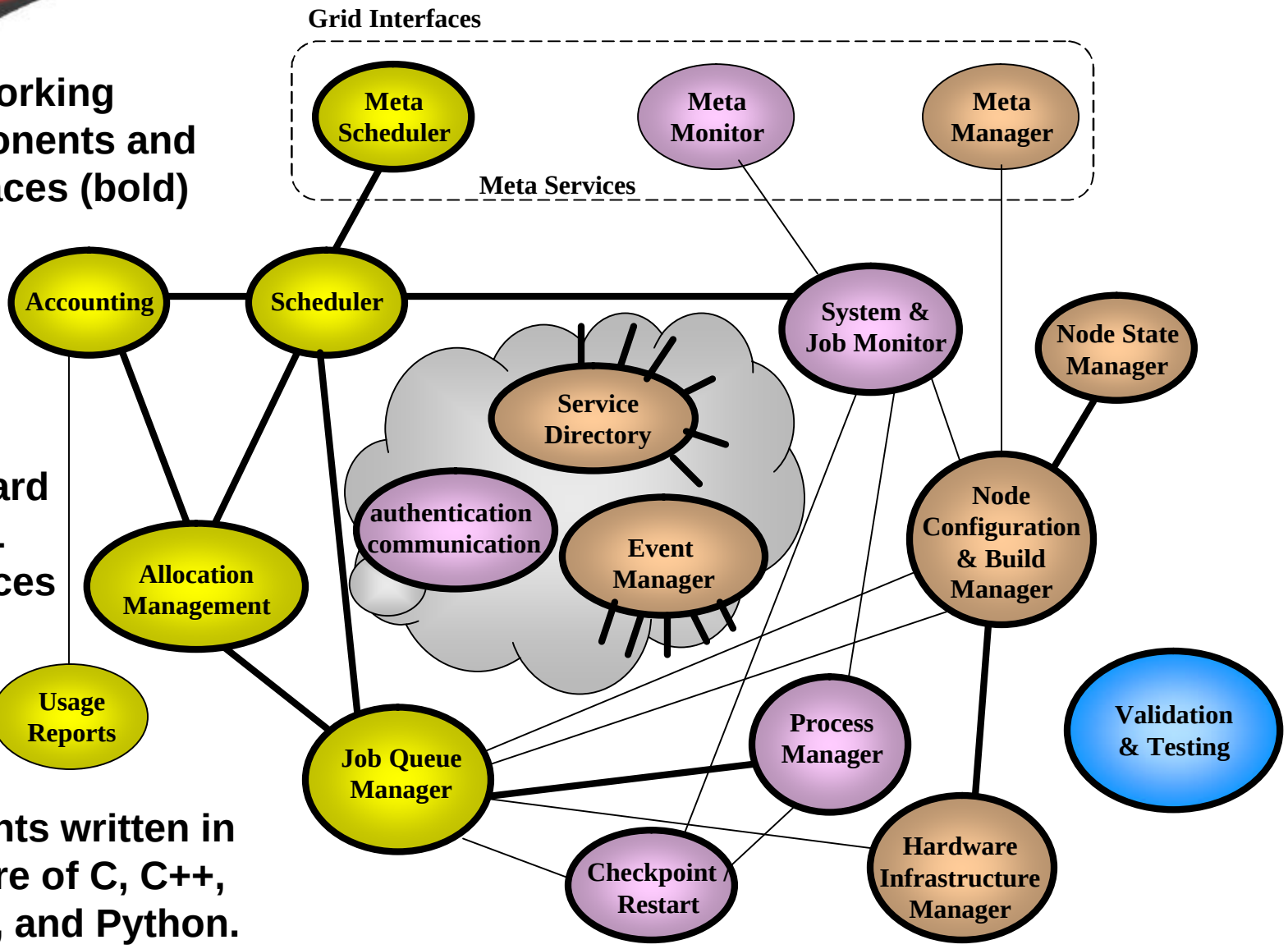
Process Manager *R.Lusk*

Checkpoint / Restart *P. Hargrove*

Process Mgmt Working Group

Progress on Integrated Suite

Working Components and Interfaces (bold)



Standard XML interfaces

Components written in any mixture of C, C++, Java, Perl, and Python.

To Learn More – Five Project Notebooks

A main notebook for general information

And individual notebooks for each working group

- Allows groups to keep track of other groups progress and comment on the items of overlap
- Allows Center members and interested parties to see what is being defined and implemented

Poster Session

Get to all notebooks through main web site
www.scidac.org/ScalableSystems

Click on side bar or at “project notebooks” at bottom of page